AUTOMATIC MULTIVARIATE NORMAL STEP-WISE REGRESSION ANALYSIS

RICHARD M. THATCHER

AUTOMATIC MULTIVARIATE NORMAL STEP-WISE REGRESSION ANALYSIS

* * * * *

RICHARD M. THATCHER

AUTOMATIC MULTIVARIATE NORMAL STEP-WISE REGRESSION ANALYSIS

by

Richard M. Thatcher

Submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE
with major in
Mathematics

United States Naval Postgraduate School
Monterey, California

1963

AUTOMATIC MULTIVARIATE NORMAL STEP-WISE REGRESSION ANALYSIS

by

Richard M. Thatcher

This work is accepted as fulfilling
the thesis requirements for the degree of
MASTER OF SCIENCE
with major in
Mathematics
from the
United States Naval Postgraduate School

ABSTRACT

We assume that p random variables, $y_1,\ldots,y_p$, are distributed according to some multivariate normal distribution (called the p variate normal). Methods of predicting the value of one, say, y_p, given the values of the other p-1 variables are discussed. A study is made of the problems encountered whenever one tries to reduce the number of variables used to predict y_p and at the same time minimize loss in prediction accuracy. Modifications of the step-wise procedure of adding predictor variables one at a time are considered in some detail, and methods of using an automatic high speed electronic computer to perform the numerous calculations involved are described. A high speed computer program was written to generate samples from any specified p variate normal.

I wish to express my sincere gratitude to Professor Jack R. Borsting, who in class introduced me to many of the mathematical concepts used in this paper, and as faculty adviser provided the guidance necessary to apply these concepts; and to Mrs. Bette Joe, for her most capable typing of this paper.

TABLE OF CONTENTS

Chapter I

INTRODUCTION

The multivariate normal distribution with p variables, referred to here as "the p variate normal" has been found to be useful as a model for a wide variety of real world phenomena. This distribution has been studied intensely in the literature and has many "nice" mathematical properties.

One of the p variate normal's most useful properties is the fact that when q of the variables are fixed, the remaining p-q variables become a p-q variate normal, which has the same variance-covariance matrix regardless of the actual fixed values of the first q variables. Where q equals p-1 the variable whose value is not fixed, say y_p, becomes a conditional normal random variable whose variance is less than the variance of y_p when the variables $y_1,\ldots,y_{p-1}$ are not fixed.

In chapter II, methods of "predicting" y_p from known fixed values of the other p-1 variables are described, and methods of measuring the accuracy of prediction in terms of variance of y_p are given. These methods require that the p variate normal be specified completely by a mean vector, U, and a variance-covariance (V-C) matrix, $\sum$. In chapter III, methods of approximating the work of chapter II using sample estimates of U and $\sum$ are described. These ideas are illustrated by an example in chapter IV.

After mastering the technique of regressing p-1 variables to form a prediction equation for the last one, y_p, we turn to the problem of eliminating variables that may not be useful in predicting the value of y_p. Variables are eliminated by removing all reference to them before the prediction equation is computed. Reasons for reducing the number of variables in regression are presented in chapter V. Later in chapter V, the process of eliminating variables from regression is illustrated by an example using a specified five-variate normal.

At present, the only known way to find the "optimum set" of r ($r \leq p-1$) variables is to compute all $\binom{p-1}{r}$ regressions. Obviously this involves extremely large numbers of computations for large p, so that methods involving fewer computations are normally used. Generally these faster methods produce "good" combinations of variables in regression but often they are not the "optimum" combination for the same number of variables in regression.

Chapters VI through IX discuss methods of searching for a satisfactorily small set of variables in regression that will reduce the conditional variance of y_p to a satisfactory level. The stepwise procedure, described in chapter VI, provides the basic procedure under study throughout the rest of the paper. Basically, this procedure consists of adding variables to regression in steps. At each step, the variable to be added is selected because its contribution to variance reduction is greatest at this step. That this procedure does not always produce optimal combinations of variables in regression is demonstrated.

Also in chapter VI a statistical test to be applied at each step when a sample is being studied is described. This test provides a criterion for halting the step-wise process which is a function of sample size, n.

In chapter VII automatic regression analysis performance by a high speed digital computer is discussed. Additional halting criteria and other improvements to the step-wise procedure are suggested. Halt criteria proposed by Miller [7] and Efroymson [3] are reviewed in light of automatic regression analysis requirements. A modification to the step-wise procedure reflecting differences in cost of observation of variables is considered.

In chapter VIII computer programs MV REGRESSION and MV SIM, written by the author, are presented. Basically, MV SIM generates samples of a specified size, n, from a given p variate normal with which MV REGRESSION performs regression analyses. MV SIM also computes regression parameters of the given p variate normal, the results of which may be used as standard for comparison purposes with results of regression analysis of the samples.

In chapter IX current and proposed studies using these high speed computer programs are outlined.

Appendix A describes the operation of program MV SIM in detail and some background on the techniques used by MV SIM to generate samples from specified p variate normals.

Appendix B describes statistical tests performed by MV SIM on sample vectors, Z, and sample (V-C) matrices, S. Results of tests performed on a number of generated samples of different sizes of a five variate normal and an 18 variate normal are given.

Chapter II

THE P VARIATE NORMAL DISTRIBUTION

In this chapter we introduce the multivariate normal distribution with p variables, hereinafter called the "p variate normal". The basic theory associated with the p variate normal is given in detail by Graybill [4] and Anderson [1]. Certain theorems and formulas that are important for later work on regression analysis are given here.

A p variate normal is completely defined by any specified pxl vector of means, U, and any pxp positive definite symmetric variance - covariance (V-C) matrix, Σ. Let:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \quad U = \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix},$$

The joint density function of the p variate normal, Y, is given by:

$$(2.1) \quad f(y_1,\ldots,y_p) = \frac{1}{(2\pi)^{(p/2)} \cdot |\Sigma|^{(1/2)}} \, e^{-1/2(Y-U)^T \Sigma^{-1} (Y-U)}$$

for $-\infty < y_i < \infty$, $i = 1,\ldots,p$.

The element σ_{ij} of Σ is the covariance between variables y_i and y_j, and u_i of U is the mean of y_i.

If the pxl vector Y is partitioned into two subvectors such that:

$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, (vectors Y_1 and Y_2 are $(p-q)x1$ and $qx1$ respectively,

$q < p$),

and if:

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

are the corresponding partitions of U and Σ, then it can be shown, [4] section 3.6, that the conditional distribution of the qxl vector Y_2 given the vector $Y_1 = Y_1{}^*$ (a constant vector), $Y_2|Y_1{}^*$, is the multivariate normal distribution with qxl mean vector

$$U_2 + \Sigma_{21}\, \Sigma_{11}^{-1}\, (Y_1{}^* - U_1),$$

and qxq V-C matrix

$$\Sigma_{22} - \Sigma_{21}\, \Sigma_{11}^{-1}\, \Sigma_{12}.$$

From the latter matrix we see the important fact that the covariance matrix of the conditional random vector $Y_2|Y_1{}^*$ does not depend upon the value of $Y_1{}^*$.

We shall represent the qxq V-C matrix of $Y_2|Y_1{}^*$ as:

(2.2) $$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\, \Sigma_{11}^{-1}\, \Sigma_{12} \,.$$

In particular, each element, $\sigma_{ii.1,\ldots,p-q}$, of this matrix ($i = p-q+1,\ldots,p$) is the conditional variance of variable y_i in Y_2, i.e. the variance of y_i when the p-q variables in Y_1 are fixed. The element σ_{ii} in the specified V-C matrix, Σ, is the variance of y_i in the original p variate normal distribution. That σ_{ii} is greater than or equal to $\sigma_{ii.1,\ldots,p-q}$ follows from formula 2.2 above, and the fact that $\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$ is positive definite. In fact, the following relationship holds (where $0 \leq R_i \leq 1$):

$$\sigma_{ii.1,\ldots,p-q} = (1-R_i^2)\, \sigma_{ii} .$$

In this formula R_i is the multiple correlation coefficient between variable y_i and vector Y_2; see [4] section 3.6.

In this paper we will consider only the case where $q = 1$. Now $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where Y is still px1, Y_1 is (p-1)x1, and Y_2 is the variable y_p. Similarly, we partition $Y^* = \begin{pmatrix} Y_1^* \\ Y_2 \end{pmatrix}$, U, and Σ so that elements Y_2, U_2, and Σ_{22} become y_p, u_p, and σ_{pp} respectively.

It follows from earlier discussions that the distribution of $y_p | Y_1^*$ is the univariate normal distribution with (scalar) mean:

$$(2.3) \qquad u_{p.1,\ldots,p-1} = U_2 + \Sigma_{21} \Sigma_{11}^{-1} (Y_1^* - U_1)$$

$$= u_p + \Sigma_{21} \Sigma_{11}^{-1} (Y_1^* - U_1);$$

and scalar variance:

$$(2.4) \qquad \sigma_{pp.1,\ldots,p-1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$$

$$= \sigma_{pp} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} .$$

Let β be the (p-1)x1 vector $\begin{pmatrix}\beta_1\\ \vdots\\ \beta_{p-1}\end{pmatrix} = (\Sigma_{21}\ \Sigma_{11}^{-1})^T$. From formula 2.3, we can write:

$$(2.5)\quad y_p|Y_1^* = u_p + \beta^T(Y_1^* - U_1) + e$$

$$= u_p + \sum_{i=1}^{p-1} \beta_i\ (y_i^* - u_i) + e$$

$$= u_p - \sum_{i=1}^{p-1} \beta_i\ u_i + \sum_{i=1}^{p-1} \beta_i\ y_i^* + e,$$

where U_1 is still $\begin{pmatrix}u_1\\ \vdots\\ u_{p-1}\end{pmatrix}$, $Y_1^* = \begin{pmatrix}y_1^*\\ \vdots\\ y_{p-1}^*\end{pmatrix}$, and e is a normally distributed random variable with mean zero. The variance of e is $\sigma_{pp.1,\dots,p-1}$, the value of which is independent of the actual values of $y_1^*,\dots,y_{p-1}^*$.

We define formula 2.5 as the <u>prediction equation for y_p associated with the p variate normal</u>. Most often we shall use it in the form:

$$(2.6)\quad y_p|Y_1^* - e = E(y_p|Y_1^*) = u_p + \sum_{i=1}^{p-1} \beta_i\cdot(y_i^* - u_i).$$

Now, if we know the fixed values of Y_1^* (in addition to U and Σ) we can use 2.6 to compute the mean of the conditional random variable $y_p|Y_1^*$. A measure of the "error" involved in using the results of 2.6 to "predict" the value of y_p, when Y_1^* is known, is given by $\sigma_{pp.1,\ldots,p-1}$. By comparison, if the value of Y_1^* is not known, one might use the original mean of y_p, u_p, to "predict" the value of y_p. The corresponding "error" of this prediction is given by σ_{pp}, which is greater than $\sigma_{pp.1,\ldots,p-1}$. The values of the scalars, β_1, in vector β are called partial regression coefficients.

Suppose the computed values of some of the partial regression coefficients β_j, β_k, etc... are zero, or close to zero. Then, obviously, insofar as estimating y_p is concerned, one can save the effort and cost of observing the values of y_j, y_k.

It often happens, especially when the number of variables, p, is large, that some of the variables, themselves, can be predicted rather accurately by a linear combination of other variables. This shows that even if none of the partial regression coefficients are close to zero, it may be possible to observe only a select few of the variables and still predict y_p nearly as accurately as when all of the variables are used.

Of course, the values of the partial regression coefficients to be used with each variable depend upon which other variables are used in combination to predict y_p. Throughout this paper, any combination of the original p-1 variables that are used to predict y_p in the manner just described will be said to be "in regression". The variables whose values are not to be used to predict y_p we shall say are "not in regression".

Once a combination of variables to be in regression have been chosen, a modified mean vector U' and V-C matrix Σ' are formed from the original U and Σ respectively by removing the u_j from U and

σ_{ij} and σ_{jk} (for all i and k) from Σ for each variable y_j that is to be "not in regression". (If q of the p-1 variables $y_1,\ldots,y_{p-1}$ are to be "not in regression", then U' is $(p-q)\times 1$ and Σ' is $(p-q)\times(p-q)$). Thus, we see that all reference to those variables not in regression is completely removed and a new p-q variate normal is defined by U' and Σ', from which new prediction equations (2.5 or 2.6) can be computed. Note that it is possible to make up $\sum_{j=1}^{p-1} \binom{p-1}{j}$ prediction equations for predicting variable y_p, one for each possible combination of variables $y_1,\ldots,y_{p-1}$.

In chapters which follow we will discuss methods of estimating the partial regression coefficients, β_i, and $\sigma_{pp.1,\ldots,q}$, etc., when the values of U and Σ are not known. Methods of choosing which variables are "best" to use in regression will be discussed. We shall also consider the problem of specifying relative "cost" of observation per unit reduction in $\sigma_{pp.1,\ldots,q}$.

Chapter III

STATISTICAL ANALYSIS OF THE P VARIATE NORMAL

In this chapter we assume that Y has a p-variate normal distribution with unknown mean vector U and V-C matrix Σ. We are now concerned with methods by which an experimenter can estimate U and Σ, and subsequently, other parameters, such as regression coefficients for prediction equations for predicting y_p; and $\sigma_{pp.1,\ldots,q}$, the conditional variance of $y_p|Y_2^*$ when variables $y_1,\ldots,y_q$ are in regression. In order to distinquish estimates of parameters from their associated theoretical values, it is convenient to develop new notation to be used throughout this paper, listed here for easy reference:

TABLE I

Notation for Theoretical Values	Meaning of Parameter	Notation for Associated Estimated Parameters
U	pxl mean vector of the p variate normal	Z
Σ	pxp V-C matrix of the p variate normal	S
β (beta)	qxl vector of regression coefficients associated with q vectors in regression	B
σ_{pp}	The element in row p, column p of Σ, which is the (unconditional) variance of y_p. y_p is arbitrarily chosen to be the variable to be predicted.	s_{pp}
$\sigma_{pp.1,\ldots,q}$	The conditional variance of $y_p|Y_1^*$, where Y_1^* is a qxl vector of fixed values of $y_1,\ldots,y_q$	$s_{pp.1,\ldots,q}$

A sample of size n can be arranged into nxp matrix form as follows:

$$\left\{\begin{matrix} y_{11}, & y_{12}, & \dots, & y_{1p} \\ y_{21}, & y_{22}, & \dots, & y_{2p} \\ \vdots & & y_{ji} & \vdots \\ y_{n1}, & y_{n2} & \dots & y_{np} \end{matrix}\right]$$

where y_{ji} represents the j th observation of variable y_i. Note that for this sample, observations of y_p, the variable later to be predicted, are also required. Sample means are computed as:

$$\bar{y}_i = \frac{\sum_{j=1}^{n} y_{ji}}{n}, \text{ for } i = 1, 2, \dots, p.$$

Sample covariances as:

$$s_{ik} = \frac{\sum_{j=1}^{n} (y_{ji} - \bar{y}_i)(y_{jk} - \bar{y}_k)}{n - 1},$$

for i, k = 1, ..., p.

For i = k the sample covariances become the sample variances:

$$s_{ii} = \frac{\sum_{j=1}^{n} (y_{ji} - \bar{y}_i)^2}{n - 1}$$

By analogy to the mean vector U and V-C matrix, Σ, we form the px1 sample mean vector, Z, and sample V-C matrix, S, as follows:

$$Z = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_p \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix}; \quad S = \{s_{ik}\} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ \vdots & & & \vdots \\ s_{p1} & & \dots & s_{pp} \end{pmatrix}.$$

It can be verified easily that Z and S are unbiased estimates of U and $\sum$ respectively; and that Z and $\left[(n-1)/n\right] \cdot S$ are maximum likelihood estimates of U and $\sum$.

To develop estimates of the parameters of the conditional distribution of $y_p | Y_1^*$ we recall that the random variable $y_p | Y_1^*$ is normally distributed with mean and variance given by equations 2.3 and 2.4. We partition Y, Y^*, Z, S, as we did Y, Y^*, U, and $\sum$, respectively in Chapter II:

$$Y = \begin{pmatrix} Y_1 \\ y_p \end{pmatrix}, \; Y^* = \begin{pmatrix} Y_1^* \\ y_p \end{pmatrix}, \; Z = \begin{pmatrix} Z_1 \\ z_p \end{pmatrix}, \text{ and } S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & s_{pp} \end{pmatrix}$$

where, as before:

$$Y_1 = \begin{pmatrix} y_1 \\ \vdots \\ y_{p-1} \end{pmatrix}, \; Y_1^* = \begin{pmatrix} y_1^* \\ \vdots \\ y_{p-1}^* \end{pmatrix} \text{ (constant vector)}$$

and,

$$Z = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_{p-1} \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_{p-1} \end{pmatrix}, \; S_{11} = \begin{pmatrix} s_{1\,1} & \cdots & s_{1\;p-1} \\ \vdots & & \vdots \\ s_{p-1\;1} & \cdots & s_{p-1\;p-1} \end{pmatrix}, \; S_{12} = (S_{21})^T = \begin{pmatrix} s_{1\;p} \\ \vdots \\ s_{p-1\;p} \end{pmatrix}.$$

Since z_i and $\frac{(n-1)}{n} \cdot s_{i\,j}$ are maximum likelihood estimates of u_i and $\sigma_{i\,j}$ respectively, for i, j = 1, ..., p, it follows from the invariant property of maximum likelihood estimates that:

$$z_p + S_{21}\, S_{11}^{-1}\, (Y_1^* - Z_1), \text{ and } \left[s_{pp} - S_{21}\, S_{11}^{-1}\, S_{12}\right] \cdot \frac{n-1}{n}$$

are maximum likelihood estimates of $u_p + \sum_{21} \sum_{11}^{-1} (Y_1^* - u_1)$, and $\sigma_{pp} - \sum_{21} \sum_{11}^{-1} \sum_{12}$ respectively.

Let:

(3.1) $$z_{p.1,\ldots,p-1} = z_p + S_{21}\, S_{11}^{-1}\, (Y_1^* - Z_1)$$

and,

(3.2) $$s_{pp.1,\ldots,p-1} = s_{pp} - S_{21}\, S_{11}^{-1}\, S_{12} \cdot \frac{n-1}{n-p-1} ,$$

It can be shown that $z_{p.1,\ldots,p-1}$ and $s_{pp.1,\ldots,p-1}$ are unbiased estimates of $u_{p.1,\ldots,p-1}$, and $\sigma_{pp.1,\ldots,p-1}$, the mean and variance of the conditional random variable $y_p|Y_1{}^*$ respectively.

Similarly, if we let $B = \begin{pmatrix} b_1 \\ \vdots \\ b_{p-1} \end{pmatrix} = (S_{21}\, S_{11}^{-1})^T$, B is a maximum likelihood estimator of $\beta = (\Sigma_{21}\, \Sigma_{11}^{-1})^T$; that is, b_i is a M.L.E. of β_i for $i = 1,\ldots,p-1$. It can be shown that B is also an unbiased estimator of β.

We can write B in the form (since S_{11}^{-1} is positive definite):

(3.3) $$B = S_{11}^{-1}\, S_{12};$$

$$\begin{pmatrix} b_1 \\ \vdots \\ b_{p-1} \end{pmatrix} = \begin{pmatrix} s_{1\ 1} & \cdots & s_{1\ p-1} \\ \vdots & & \vdots \\ s_{p-1\ 1} & \cdots & s_{p-1\ p-1} \end{pmatrix} \cdot \begin{pmatrix} s_{1p} \\ \vdots \\ s_{p-1\ p} \end{pmatrix},$$

Equations 3.3 are called the normal equations.

Substituting $z_{p.1,\ldots,p-1}$ for $u_{p.1,\ldots,p-1}$, we obtain an unbiased estimate for the value of the prediction equation, 2.6, by:

(3.4) $$\widehat{E\,(Y_p|Y_1{}^*)} = z_{p.1,\ldots,p-1}$$

$$= z_p + S_{21}\, S_{11}^{-1}\, (Y_1^* - Z_1)$$

$$= z_p + \sum_{i=1}^{p-1} b_i\, (y_i^* - z_i).$$

Chapter IV

AN EXAMPLE

Assume that an experimenter wishes to gather data from some process involving five variables, which he assumes to be related according to a five variate normal distribution. Suppose this five variate normal actually is defined (completely) by the following theoretical vector, U, and V-C matrix, Σ :[1]

$$(4.1)\qquad U = \left\{\begin{array}{r} 7.4600 \\ 48.1500 \\ 11.7700 \\ 30.0000 \\ 95.4200 \end{array}\right\} = \left\{\begin{array}{c} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{array}\right\},$$

$\frac{1}{n} \Sigma x_i$ [illegible]

(4.2)

$$\Sigma = \left\{\begin{array}{c|rrrrr} & x_1 & x_2 & x_3 & x_4 & y \\ \hline x_1 & 34.6025 & 20.9233 & -31.0517 & -24.1667 & 64.6633 \\ x_2 & 20.9233 & 242.1408 & -13.8783 & -253.4167 & 191.0792 \\ x_3 & -31.0517 & -13.8783 & 41.0258 & 3.1667 & -51.5192 \\ x_4 & -24.1667 & -253.4167 & 3.1667 & 280.1667 & -206.8083 \\ y & 64.6633 & 191.0792 & -51.5192 & -206.8083 & 226.3133 \end{array}\right\}.$$

Using developments of chapter II, we let $Y = \left\{\begin{array}{c} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{array}\right\}.$

[1] The value of U and Σ used here were computed as sample vector Z, and V-C matrix S using data from table 20.4, page 647 of HALD [H]. The results of tables 20.5 and 20.6 of Hald were used to verify the results of computer program MVSIM, which performed most of the computations required for this paper.

We know that we are going to be given values of y_1, y_2, y_3, y_4, from which we will predict y_5. Hence, we must set up the prediction equation for y_5. (equation 2.6). Accordingly, we partition U and Σ as:

$$U = \begin{bmatrix} \begin{Bmatrix} 7.4600 \\ 48.1500 \\ 11.7700 \\ 30.0000 \end{Bmatrix} \\ \\ (95.4200) \end{bmatrix} = \begin{bmatrix} U_1 \\ \\ \\ U_2 \end{bmatrix}$$

$$\Sigma = \left[\begin{array}{cccc|c} 34.6025 & 20.9233 & -31.0517 & -24.1667 & 64.6633 \\ 20.9233 & 242.1408 & -13.8783 & -253.4167 & 191.0792 \\ -31.0517 & -13.8783 & 41.0258 & 3.1667 & -51.5192 \\ -24.1667 & -253.4167 & 3.1667 & 280.1667 & -206.8083 \\ \hline 64.6633 & 191.0792 & -51.5192 & -206.8083 & 226.3133 \end{array}\right]$$

$$= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

For β, we get

$$\beta = \left(\Sigma_{21}\, \Sigma_{11}^{-1}\right)^{T} = \begin{Bmatrix} 1.5513 \\ .5103 \\ .1021 \\ -.1438 \end{Bmatrix} = \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{Bmatrix}.$$

The prediction equation for y_5 becomes:

$$(4.3) \quad E\,(y_5|Y_1{}^*) = u_5 - \sum_{i=1}^{4} \beta_i u_i + \sum_{i=1}^{4} \beta_i y_i{}^*$$

$$= 95.4200 - (1.5513 \quad .5103 \quad .1021 \quad -.1438)\cdot\begin{Bmatrix} 7.4600 \\ 48.1500 \\ 11.7700 \\ 30.0000 \end{Bmatrix}$$

$$+ (1.5513 \quad .5103 \quad .1021 \quad -.1438)\cdot\begin{Bmatrix} y_1{}^* \\ y_2{}^* \\ y_3{}^* \\ y_4{}^* \end{Bmatrix},$$

or,

(4.4)

$$\overline{E\,(y_5|Y_1{}^*)} = 62.3881 + 1.5513\, y_1{}^* + .5103\, y_2{}^* + .1021\, y_3{}^* - .1438\, y_4{}^*.$$

The variance of $y_5|Y_1{}^*$, $\sigma_{55.1,2,3,4}$, (the conditional variance of y_5 given y_1, y_2, y_3, y_4,), is:

$$\sigma_{55.1,2,3,4} = \Sigma_{22} - \Sigma_{21}\,\Sigma_{11}^{-1}\,\Sigma_{12}$$

$$= 226.3133 - 222.3242 = 3.9891$$

which is a measure of the prediction error when formula 4.4 is used to predict y_5 when $Y_1{}^*$ is known. By comparison, if the values of $Y_1{}^*$ were ignored and if, instead, the value $u_5 = E\,(y_5) = 95.4200$ was always used to estimate y_5, the corresponding measure of the prediction error would be $\sigma_{55} = 226.3133$.

Thus, by knowing the mean vector U, and the V-C matrix $\sum$, as given by formulas 4.1 and 4.2, we can set up the above prediction equation, 4.4. Then for any set of values y_1, y_2, y_3, y_4, we can make an accurate prediction of y_5 without observing its value.

The problem facing the experimenter is more complicated than the one discussed in the preceding paragraphs. This is because he does not know the values of mean vector, U and the V-C matrix, $\sum$. All he knows is that (by assumption) y_1, y_2, y_3, y_4, y_5 are distributed according to some five variate normal distribution, and, therefore, are completely specified by some theoretical mean vector U and V-C matrix $\sum$ whose actual values he will never know.

Assume the experimenter draws a sample of size 500 from this five variate normal distribution (specified by equations 4.1 and 4.2). He then computes all sample means, variances, and covariances ($\bar{y}_i = z_i$, s_{ii}, s_{ij} respectively) and forms the sample mean vector, Z, and sample V-C matrix S as defined in chapter III. Suppose, as an example, he obtains the following results upon drawing a sample of size 500:

$$Z = \begin{bmatrix} \begin{Bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{Bmatrix} \\ \\ (z_5) \end{bmatrix} = \begin{bmatrix} \begin{Bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \bar{y}_4 \end{Bmatrix} \\ \\ (\bar{y}_5) \end{bmatrix} = \begin{bmatrix} \begin{Bmatrix} 7.7764 \\ 48.7155 \\ 11.5687 \\ 29.3039 \end{Bmatrix} \\ \\ (96.4816) \end{bmatrix}$$

$$S = \{s_{ij}\} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

$$= \left[\begin{array}{cc} \begin{Bmatrix} 34.6160 & 20.1495 & -32.8248 & -21.5656 \\ 20.1495 & 217.9056 & -18.3443 & -223.1718 \\ -32.8248 & -18.3443 & 42.8371 & 7.4936 \\ -21.5656 & -223.1718 & 7.4936 & 242.3209 \end{Bmatrix} & \begin{Bmatrix} 64.0139 \\ 171.9229 \\ -57.4059 \\ -180.4153 \end{Bmatrix} \\ \begin{pmatrix} 64.0139 & 171.9229 & -57.4059 & -180.4153 \end{pmatrix} & \begin{pmatrix} 211.0392 \end{pmatrix} \end{array}\right].$$

To estimate β by B we compute

$$B = \left[S_{21}\ S_{11}^{-1}\right]^T = \begin{Bmatrix} 1.6360 \\ .5921 \\ .1774 \\ -.0590 \end{Bmatrix}$$

Hence, the estimate of the prediction equation 3.4 becomes:

$$\widehat{E(y_5|Y_1^*)} = z_5 - \sum_{i=1}^{4} b_i z_i + \sum_{i=1}^{4} b_i y_i^*$$

$$= 54.5921 + 1.6360\, y_1^* + .5921\, y_2^* + .1774\, y_3^* - .0590\, y_4^*.$$

The unbiased estimate of the conditional variance of $y_5|Y_1^*$ would be:

$$s_{55\cdot 1,2,3,4} = \left[s_{55} - S_{21}\ S_{11}^{-1}\ S_{12} \right] \cdot \frac{n-1}{n-p-1}$$

$$= 4.0368 \times \frac{499}{495} = 4.0694$$

Now the experimenter is in a position to predict the value of y_5

given a set of values y_1^*, y_2^*, y_3^*, y_4^*. For example, suppose he is given that the $y_i^* = u_i$, the true means of the y_i. (Of course, he doesn't know that these are true means).

Using the true prediction equation, we get (see 4.3):

$$E\ (y_5 | Y_1^* = U) = u_5 - \sum_{i=1}^{4} \beta_i u_i + \sum_{i=1}^{4} \beta_i u_i = 95.4194.$$

The experimenter would estimate this value as:

$$\begin{aligned}\widehat{E\ (y_5 | Y_1^* = U)} = 54.5921 &+ (\ \ 1.6360) \cdot (\ 7.4600) \\ &+ (\ \ .5921) \cdot (48.1500) \\ &+ (\ \ .1774) \cdot (11.7700) \\ &+ (-\ .0590) \cdot (30.0000) = 95.6243\ .\end{aligned}$$

Chapter V

REDUCTION IN THE NUMBER OF VARIABLES IN REGRESSION - INTRODUCTION

Experience has shown that when the number of variables, p, is large, say over 20, usually a relatively small number of variables can be found to use in regression to predict y_p nearly as accurately as when all p-1 variables are used in regression, [9], page 20. Finding such a small combination of variables is desirable for a number of reasons:

1) The prediction equation, has fewer terms; thus it is easier to compute a predicted value of y_p.

2) Fewer variables need to be observed in order to make a prediction of y_p. Presumably this would result in reducing the cost of observing variables for each prediction of y_p.

3) When p is large, the prediction equation involving p-1 variables requires many computations. Step-wise procedures, described later, when yielding a relatively small number of variables in regression produce a prediction equation with much less effort.

4) When the regression is being performed on a sample, variables that do not contribute much variance reduction of y_p can actually cause the prediction equation to yield a worse fit to the underlying (specified) p variate normal than would result if they were omitted from regression. The reason is that the longer equation

can overfit the sample and ascribe some of the variation due to small scale random fluctuations to one of the predictors "by accident".

As one would suspect, whenever a single variable, y_k, is added to regression the old conditional variance, say, $\sigma_{pp.1,\ldots,q}$ is always greater than or equal to the new conditional variance, $\sigma_{pp.1,\ldots,q,k}$. However, usually the amount by which $\sigma_{pp.1,\ldots,q}$ is reduced becomes small as the number of variables in regression increases, even though optimal combinations for each number of variables in regression are used. To illustrate this idea, let us consider an example of a regression problem under ideal conditions. That is, we shall examine a p variate normal specified in terms of vector U and matrix $\sum$.

We first compute the prediction equation for y_p, and the associated conditional variance $\sigma_{pp.1,\ldots,q}$, for each possible combination of variables $y_1,\ldots,y_{p-1}$ in regression $\left(\sum_{j=1}^{p-1} \binom{p-1}{j}\right.$ sets of prediction equations to solve). We shall then group the results according to number of variables in regression, and from each group pick the "optimal" combination of variables in regression; that is, the combination of variables, say, $y_1,\ldots,y_q$, in regression producing the smallest $\sigma_{pp.1,\ldots,q}$.[1]

[1] Clarification of notation: The reader should understand that whenever a "combination of variables in regression, say, $y_1,\ldots,y_q$" and the associated conditional variance, "$\sigma_{pp.1,\ldots,q}$, is discussed as in the preceding paragraph, the q variables in regression are not necessarily meant to be regarded as the first q variables as defined by position in the original vector U and matrix $\sum$. In other words, in order to ease notational difficulty, variables in regression are temporarily relabeled $y_1,\ldots,y_q$.

With this grouping, we can now start with one variable in regression and add to the number of variables in regression one at a time, each time choosing the "optimal" combination of variables for that group, until we decide that adding more variables to regression will not reduce $\sigma_{pp.1,\dots,q}$ enough to make it worthwhile.

In our example, we shall use the five variate normal as defined by 4.1 and 4.2. To compute the prediction equation using only y_1 in regression, the prediction equation becomes:

$$E\,(y_5|Y_1^*) = u_5 - \beta_1\, u_1 + \beta_1\, y_1^*$$

$$\text{where } \beta_1 = \beta = \left[\Sigma_{21}\,\Sigma_{11}^{-1}\right]^T = 64.6633 \cdot \frac{1}{34.6025} = 1.8687,$$

$$\text{and } \sigma_{55\cdot 1} = \Sigma_{22} - \Sigma_{21}\,\Sigma_{11}^{-1}\,\Sigma_{12}$$

$$= 226.3133 - \frac{64.6633 \times 64.6633}{34.6025} = 105.4739\ .$$

Similarly, we compute partial regression coefficients, β_i, for all 15 possible combinations of the variables y_1, y_2, y_3, y_4 in regression. Table II shows the results.

Table II

Variables in Regression	q	β_1	β_2	β_3	β_4	Associated Conditional Variance $\sigma_{55\cdot 1,\ldots,q}$
y_1	1	1.8687				105.4739
y_2	1		.7891			75.5280
y_3	1			- 1.2557		161.6167
* y_4	1				- .7381	73.6553 ←
*y_1 y_2	2	1.4683	.6622			4.8261 *
y_1 y_3	2	2.3125		.4945		102.2551
y_1 y_4	2	1.4399			- .6139	6.2303 ←
y_2 y_3	2		.7313	- 1.0083		34.6208
y_2 y_4	2		.3108		- .4569	72.4065
y_3 y_4	2			- 1.1998	- .7245	14.6447
y_1 y_2 y_3	3	1.6959	.6569	.2500		4.0096
*y_1 y_2 y_4	3	1.4519	.4160		- .2365	3.9982 ←
y_1 y_3 y_4	3	1.0518		- .4100	- .6427	4.2368
y_2 y_3 y_4	3		- .9234	- 1.4479	- 1.5570	6.1506
*y_1 y_2 y_3 y_4	4	1.5513	.5103	.1021	- .1438	3.9891

Note: Each group is identified by the value of q.

* Indicates the "optimal" combination of variables for the group (for that number of variables in regression).

We now select the optimal combination from each group of variables in regression as follows (we omit the partial regression coefficients):

Table III

q Group Number	Variables in Regression (Optimal Combination)	Associated Conditional Variances of y_5
0	None	$\sigma_{55} = 226.3133$
1	y_4	$\sigma_{55 \cdot 4} = 73.6553$
2	y_1, y_2	$\sigma_{55 \cdot 1,2} = 4.8261$
3	y_1, y_2, y_4	$\sigma_{55 \cdot 1,2,4} = 3.9982$
4	y_1, y_2, y_3, y_4	$\sigma_{55 \cdot 1,2,3,4} = 3.9891$

From table III we immediately see that most of the reduction of the conditional variance of y_5 can be done by introducing only two of the possible four variables into regression, namely y_1 and y_2. Very little more is accomplished by using the other two variables, given that y_1 and y_2 are going to be used in regression.

Note that the five variate normal is easily handled by an electronic computer because only $\sum_{j=1}^{4} \binom{4}{j} = 15$ prediction equations had to be computed, none with more than five variables involved. On the other hand, the 18 variate normal, for example, requires $\sum_{j=1}^{17} \binom{17}{j} = 131{,}071$ prediction equations, most of which involve many variables. Hence this procedure is not always feasible even when today's high speed electronic computers are available.

It is interesting to note that when all four variables are in regression the values of the regression coefficients do not suggest which variables might be best to eliminate from regression. In fact, none of the values are close enough to zero to indicate that any should be removed.

In this chapter it has been shown that we can expect the amount of reduction in the conditional variance of y_p to be less per variable added to regression when the number of variables in the optimal combination becomes larger. Thus, if one were willing to state in advance his maximum allowable value of the conditional variance of y_p, the problem would be a straight forward one of searching table II for the minimum number of variables producing that conditional variance or less. We now restate this same problem in the above terms:

"To find some satisfactorily small number of variables, q, $(q \leq p-1)$, that, when used to predict y_p, reduces, $\sigma_{pp.1,\ldots,q}$, to some satisfactorily small fraction of the unconditional variance of y_p, σ_{pp}."

Chapter VI

THE STEP-WISE PROCEDURE

We now discuss an alternate procedure of searching for optimal combinations of variables in regression, called the step-wise procedure. This procedure has the advantage of reducing the number of prediction equations to be solved from $\sum_{r=1}^{p-1} \binom{p-1}{r}$, as in chapter V, to p-1 or less, thus keeping the number of computations to within the capability of today's high speed electronic computers. We shall see that the combination of variables selected by this method is not always optimal, i.e., it is possible that a different set of the same number of variables might yield a more accurate prediction equation for y_p. However, practical experience indicates that sets decidedly better than those discovered by the procedure outlined in this chapter are rare [A], page 19. We shall discuss additional problems encountered when the step-wise procedure is applied to a sample. The need for statistical tests at each step is demonstrated and an actual test is developed.

The step-wise procedure is as follows: At each step every variable not yet in regression is examined to see how much the conditional variance of y_p would be decreased if it, alone, were added to the variables already in regression, i.e., assuming q variables are already in regression, the quantity $\sigma_{pp.1,\ldots,q} - \sigma_{pp.1,\ldots,q,m}$ is computed for each variable, y_m, still not in regression. The variable to be added to regression is y_k, the variable for which this computation is greatest; i.e., y_k is chosen from the variables not in regression, y_m, so that

$$\sigma_{pp.1,\ldots,q} - \sigma_{pp.1,\ldots,q,k} = \max_m \left[\sigma_{pp.1,\ldots,q} - \sigma_{pp.1,\ldots,q,m}\right]$$

or equivalently, such that:

$$\sigma_{pp.1,\ldots,q,k} = \min_m \left[\sigma_{pp.1,\ldots,q,m}\right].$$

We illustrate this procedure by applying it to the p variate normal specified by equations 4.1 and 4.2. This illustration can be followed most easily if reference is made to table II of chapter V:

Step I: Compute all four conditional variances of group 1, and choose the smallest value (73.6553).

action: add variable y_4 to regression

results: variables in regression: y_4

$$\sigma_{55\cdot4} = 73.6553$$

Step II: Compute the conditional variances of group 2 that include variable y_4 in regression, and choose the smallest value (6.2303).

action: add variable y_1 to regression

results: variables in regression: y_1, y_4

$$\sigma_{55\cdot1,4} = 6.2303$$

Step III: Compute the conditional variances of group 3 that include variables y_1 and y_4 in regression, and choose the smallest value (3.9982).

action: add variable y_2 to regression

results: variables in regression: y_1, y_2, y_4

$$\sigma_{55\cdot 1,2,4} = 3.9982$$

Step IV: Add the last variable

results: variables in regression: y_1, y_2, y_3, y_4

$$\sigma_{55\cdot 1,2,3,4} = 3.9891$$

As in the preceding chapter, we immediately see that most of the conditional variances of y_5 can be eliminated by using only two of the possible four variables in regression. However, this time the pair chosen were variables y_1 and y_4 instead of y_1 and y_2; producing a conditional variance of 6.2303 instead of 4.8261.

The step-wise procedure is equally applicable to analysis of a sample of size n. In this case all information is obtained from the sample vector, Z, and sample V-C matrix, S. In particular, the values of the sample conditional variances, $s_{pp.1,\ldots,q}$, rather than $\sigma_{pp.1,\ldots,q}$ are used at each step to determine the next variable to enter regression. As before, p-1 prediction equations, and associated estimated conditional variances of y_p, $s_{pp.1,\ldots,q}$, can be obtained. Each succeeding equation will contain one more variable in regression, and usually will have a smaller value of $s_{pp.1,\ldots,q}$.[1] Now, as in chapter VI, the most acceptable combination of variables in regression, for which the estimated conditional variance of y_p is small enough, can be chosen.

[1] The exception can occur when the sample size, n, is small. See Hald's example table 20.6, where n = 13. [H].

At this point we must consider a problem that is ever present whenever a sample is used as a source of information. In the present case the problem is stated as follows: How do we know that the sample size, n, was large enough, so that the conditional variance associated with the combination we have just selected is accurate? (We will always assume that n is greater than p).

Intuitively, if n is just a little larger than p we should not have much confidence in sample vector Z and sample V-C matrix S, nor in the estimated regression coefficients or conditional variances of y_p. In fact we shouldn't be surprised if a second sample of the same size were to produce a completely different set of variables when the same step-wise procedures are used. On the other hand, as n approaches infinity the samples Z and S approach the true values of U and Σ. It is clear that at each step, each variable that is a candidate to enter regression should be given a statistical test of some kind.

Suppose q variables, $y_1,\ldots,y_q$, are already in regression with estimated conditional variance of y_p given by $s_{pp.1,\ldots,q}$; and suppose that we are considering variable y_k for addition to regression. It can be shown that if actually $\sigma_{pp.1,\ldots,q,k} = \sigma_{pp.1,\ldots,q}$, then the statistic

$$(6.1) \qquad F = \frac{(n-q-1)\ s_{pp.1,\ldots,q} - (n-q-2)\ s_{pp.1,\ldots,q,k}}{s_{pp.1,\ldots,q,k}}$$

has the F distribution with 1 and n-q-1 degress of freedom [4], section 6.4. Furthermore, statistic F will tend to be greater than

F (1, n-q-1) if $\sigma_{pp.1,\ldots,q,k}$ is actually less than $\sigma_{pp.1,\ldots,q}$.

We immediately encounter a new complication: The above statistic F behaves as stated above so long as variable y_k is studied by itself. However, the selection of y_k from among those variables still not in regression was not completely at random. y_k was chosen at this time because it was estimated to be the "best" variable to add at this step. In other words, we are in effect computing F for a number of variables and choosing the variable for which F is the largest. It is important to realize that due to this method of selection, the F statistic used with the selected variable y_k will <u>tend</u> to be larger than would be expected on the average if variable y_k were to be studied as an individual variable alone. Intuitively, this effect should be stronger with the first variables added to regression, since those variables for which F is large due to randomness, are removed from those not in regression early. Suggested procedures for compensating for this are discussed in a later chapter.

Let α be the probability of erroneously concluding that $\sigma_{pp.1,\ldots,q,k}$, is less than $\sigma_{pp.1,\ldots,q}$ whenever actually they are equal (α is usually chosen to be .05). This error is usually called the type one error. Suppose now, at each step we compute the statistic F of formula 6.1, and compare with the value of $F_{\alpha(1,n-q-1)}$ which can be found in tables of the F distribution. If the sample size is too small the <u>power</u> of the test will be low. This means that the actual difference between $\sigma_{pp.1,\ldots,q}$ and $\sigma_{pp.1,\ldots,q,k}$, can be substantial and still, the probability that the computed statistic, F, will exceed $F_{\alpha(1,n-q-1)}$ can be small. (Of course, this probability will always be greater than α). This error is usually called the type two error.

On the other hand, given that $\sigma_{pp.1,\ldots,q}$ is actually greater than $\sigma_{pp.1,\ldots,q,k}$ (the only alternative being that they are equal) regardless of how small the actual difference, the probability that statistic F exceeds $F_{\alpha(1,n-q-1)}$ can be made as close to one as we please by increasing sample size, n, indefinitely.

Meanwhile, among those variables for which $\sigma_{pp.1,\ldots,q}$ is actually equal to $\sigma_{pp.1,\ldots,q,k}$, approximately α × 100 percent are expected to "pass" the F test (i.e., $F > F_{\alpha(1,n-q-1)}$) independently of the sample size, n.

We have just seen that the two important factors that affect the probability that variable y_k will pass a particular F test are the amount by which the actual values of $\sigma_{pp.1,\ldots,q}$, and $\sigma_{pp.1,\ldots,q,k}$ differ, and the size of the sample, n. Thus, the decision rule we might use is to terminate the step-wise procedure at any step that all variables still not in regression fail to pass the F test. With this decision rule, the F test will limit the variables in regression to those whose contribution to reduction in conditional variance of y_p appear to be large enough for the given sample to measure.

In the next chapter we shall consider additional halt criteria which an experimenter may wish to impose on the step-wise process.

Chapter VII

AUTOMATIC REGRESSION ANALYSIS - CRITERIA FOR HALTING STEP-WISE REGRESSION

In this chapter we shall develop useful procedures for conducting automatic regression analysis on a sample of size n of a p variate normal using a high speed electronic computer. Efroymson [3] has developed an algorithm very suitable for computer use in which any single variable can be added to, or eliminated from regression (depending upon its former status). At any step the regression coefficients, conditional variance of y_p, multiple correlation coefficient of y_p on the variables in regression, and many other desirable parameters can be computed easily and printed out. Useful criteria for halting the regression process are discussed and developed.

Given a sample of size n of a p variate normal, formulas for computing vector Z and matrix S have already been described. Also, basically we shall use the step-wise procedure of adding variables to regression. The most important remaining problem is to consider how the user of an automatic regression analysis computer program can specify in advance of the computer run, reasonable criteria for halting the step-wise procedure.

So far, it appears that a satisfactory criteria for stopping the regression process has never been fully developed to suit automatic step-wise regression. Miller [7] proposes adding variables until the F test fails. He also proposes a method of adjusting the level for which the critical F is chosen ($1 - \alpha$ in chapter VII) to compensate for the fact that the method of choosing each variable to enter regression is not a random choice:

> In order to derive a test for the statistical significance of X_j, the following analysis may be performed: When a predictor is chosen at <u>random</u> from a group of predictors, an F test is performed where the critical F is usually taken at the 95% level. This allows for a one in twenty chance for considering this predictor significant when in fact it is not. In the screening procedure the selection of X_j is <u>not</u> a random choice. Therefore, it is necessary to determine at what probability level the critical F should be taken while still specifying a one in twenty chance occurrence.
>
> For the screening procedure it appears proper to make the level for which the critical F is chosen a function of the number of <u>possible</u> predictors, n. The ordinary 95% level F can be expressed as
>
> $$F_{.95} = F_{(1 - 1/20)},$$
>
> and for the screening procedure the 95% level is
>
> $$F^*_{.95} = F_{(1 - 1/20 \cdot n)}.$$

Intuitively, Miller's solution seems to be somewhat extreme. For example, if p = 51 (and $\alpha = .05$) then at the first step the level chosen for the critical F, α', would be computed as follows:

$$1 - \frac{\alpha}{p} = 1 - \frac{1}{20\times50} = .998; \quad \alpha' = 1 - .998 = .001;$$

so that the value used for comparison would be $F_{.001}\ (1,49) = 12.2$, rather than $F_{.05}\ (1,49) = 4.03$ when no adjustment is made. In this case the critical F value is arbitrarily tripled only because there are 50 variables still not in regression. Granted that the critical F should be adjusted upward in order to maintain a "one in twenty chance occurrence", it would seem that due to lack of information as to the extent of this non-random effect, one should make such an adjustment more conservatively than this.

Perhaps a satisfactory "hedge" might be to use the adjusted level:

$$\alpha' = \frac{\alpha}{\log p}, \text{ or } \alpha' = \frac{\alpha}{\log p} K,$$

where K is a constant inserted by the program user for his particular sample.

Conceivably, one might wish to make no adjustment at all for this effect because the consequences of increasing the type two error during the early steps are so detrimental to the step-wise procedure.

Efroymson [3] proposes two F tests at each step. His program first compares each variable, y_i, currently in regression with an appropriate "min F" critical value to see if it still passes the F test of significance. If such a variable is discovered, the action at that step is to remove the variable from regression. By setting min F to a value slightly less than the standard critical value used for adding variables, the possibility of creating an endless loop is avoided.

This feature is appealing because new combinations in regression obtained in this manner are always more nearly optimal (as far as the sample is concered) than was the preceding combination of the same size; yet the number of computer instructions required to do this operation is minimal.

In chapter VI it was shown that the choice of $F_{\alpha(1,n-q-1)}$ as the value of critical F is made in an attempt to limit the variables in regression to those whose contribution to reduction in conditional variance of y_p is large enough for the given sample to measure. It is clear that Efroymson's double F test contributes to this effort by insuring that all variables in regression continue to pass the F test even after subsequent variables have been added.

It is impossible to anticipate here all uses for different combinations of specified stopping criteria. We already have seen that statisticians so far have only provided general guide lines in this area. This is mainly because each individual p variate normal distribution has its own set of complications, and for each computer run on a given sample the experimenter may have varying amounts of prior information regarding the p variate normal he is studying. Thus, for any automatic regression analysis computer program it is important that the user of the program be able to specify halt criteria with as much flexibility as possible.

Perhaps the most important aspect of each halt criteria is that it must be specifiable in a manner most meaningful to the experimenter. For example, some experimenters under certain conditions may not look upon the F test of chapter VI as being useful to him at all. Quite likely, he may wish to replace $F_{\alpha(1,n-q-1)}$ with a value, say λ, to be the critical amount of reduction of the variance of y_p as a stopping rule; or, he may want to specify both critical values. Although λ and $F_{\alpha(1,n-q-1)}$ are in different units, it is clear that the λ test is equivalent to an F test, so that in specifying both tests the experimenter is merely having the computer apply whichever test is the most stringent at each step.

The following example illustrates most of the points covered in the last few paragraphs. We show here how a suitable choice of min F and critical F, artificially chosen, can aid Efroymson's double F test procedure to find a more nearly optimal combination of variables in regression than already obtained by the step-wise procedure at a previous step. To do this we take an example worked out by Hald [6], section 20.3. In this example, Hald used data from a sample of size 13 of a five variate distribution which we will assume here to be normal. The sample vector, Z, and sample V-C matrix, S, are the same as those shown by equations 4.1 and 4.2 in this paper, which in chapter IV were used to define U and Σ respectively. In the following illustration we shall consider 4.1 and 4.2 to be computed Z and S as in Hald's example.

From Hald's example we compute the F statistic (6.1) for each variable in regression and not in regression at each step. See table IV below. For variables in regression, y_k, the value computed is the F statistic that would be computed for y_k if it, alone, were removed from regression first. These values pertaining to variables currently in regression are underscored in table IV. In order to illustrate the above points, the F test using $F_{\alpha(1,n-q-1)}$ was eliminated.

We now choose the artifical values of critical F and min F to be 3.5 and 3.0 respectively. With this choice we shall obtain the optimal combination of variables y_1 and y_2, where the regular forward step-wise procedure yielded variables y_1 and y_4 in chapter IV.

Table IV

Step	Variables in Regr. Before This Step	Computed F Statistic(6.1)			
		y_1	y_2	y_3	y_4
1	0	12.60	21.96	4.40	(22.80)
2	y_4	(108.16)	.17	40.30	22.80
3	y_1, y_4	108.16	(5.03)	4.24	159.21
4	y_1, y_2, y_4	154.02	5.03	.01	1.86
5	y_1, y_2	the optimal combination for two variables in regression			

The variables added to or eliminated from regression were chosen according to Efroymson's double F test procedure. Recall that no variable was to be added at any given step if the F value of one of the variables already in regression got below 3.0 (min F). Hence at step 4, variable y_4 was eliminated yielding the optimal combination y_1, y_2. At each previous step, the variable added (whose value is enclosed in parentheses) was chosen because its F statistic was the largest among those still not in regression and was also greater than the critical F value which was artifically chosen to be 3.5.

This example illustrates some complexities that arise during the regression process that are still not completely explainable analytically. For instance, the relative values of statistic F changed drastically as the combination of variables in regression changed. These values correspond to relative amounts of reduction of conditional variance that would be due to the corresponding variable if it were (or is) in regression. Thus, when y_4 was added to regression at the first step the relative contribution in variance reduction due to y_1 jumped from 13 to 108, implying that y_1 and y_4 are much more powerful together than their sum when each is used alone.

This example also suggests reasons why an experimenter may wish to specify critical F values artificially, especially if results of prior computer runs are available.

It was suggested earlier that instead of keeping track of computed values of (6.1) requiring specification of artificial critical F's on the part of the experimenter, it might be simpler for him to keep track of actual amounts of variance reduction of y_p and make up artifical values of λ in units of variance reduction of y_p. Also it is clear that the experimenter may wish to specify a value, say, $\min\lambda \leq \lambda$, which would become the critical amount of variance reduction required of each variable in regression, in order to stay in regression.

The following summary lists a few useful halt criteria which the experimenter may wish to specify before the automatic regression analysis is performed on a given sample. The automatic regression program should permit the experimenter to specify any combination of these criteria for any given computer run:

1. $F_{\alpha(1,n-q-1)}$ and min F (Chapter VI)
2. λ and $\min\lambda$ (defined above)
3. Stop when the conditional variance of y_p gets as low as

V percent of the original variance, s_{pp}.

4. Stop when the conditional variance of y_p gets as low as T.

5. Stop when W variables have been added to regression.

In chapter IX we shall propose a procedure using some of the above halt criteria in searching for an optimal combination of variables in regression.

In chapter V it was stated that one good reason for reducing the number of variables in regression might be to reduce the cost of observing the variables from which each future prediction of y_p is to be computed. Often some of the variables cost considerably more to observe than others, and the experimenter may not be so interested in reducing the total <u>number</u> of variables to observe, as he is in reducing the <u>total cost to observe</u> the values of the variables in regression for each prediction of y_p to be made later. Thus, it is desirable that the experimenter be able to specify observation costs, c_i, (say, in dollars) for each "independent" variable $y_1,\dots,y_{p-1}$, and have the automatic regression analysis operation reflect these costs when selecting variables to go into regression.

The "cost option" should differ from the regular option only in the criteria used at each step to determine which variable is to be added to regression. Recall that the regular option calls for chosing the variable that will reduce the variance of y_p the most, to be the variable added.

In the cost option, at each step, those variables still not in regression are determined. Then, instead of y_k for which $\sigma_{pp.1,\dots,q,k}$ is least, (as estimated by $s_{pp.1,\dots,q,k}$), y_j is chosen for which

$c_j / (\sigma_{pp.1,\ldots,q} - \sigma_{pp.1,\ldots,q,j})$ is least; i.e., y_j is chosen on the basis that it is cheapest in terms of "dollars" to observe per unit of variance reduction of y_p, due to adding y_j. It is clear that the standard option is just a special case of the cost option in which the observation costs are all specified to be equal.

Since optimality is now measured in terms of minimum cost to observe per unit of variance reduction instead of maximum variance reduction, the program user must be able to specify a halt criterion so that whenever the cost to observe a variable in regression becomes greater than, say, max C, the program will remove it, and whenever all variables still not in regression would cost more than, say, C dollars per unit of variance reduction, if added, the program should halt regression. Now, min λ and λ are not needed as halting criteria for the cost option. However, the experimenter should still have the option of including other halt criteria summarized above.

To summarize, neither Miller's nor Efroymson's stopping rules are optimal. Both basically use only the statistical F test of chapter VI as a decision rule for halting. It has been illustrated here that additional decision criteria that can be specified by the experimenter in terms more meaningful to him, may greatly facilitate his search for optimal combinations of variables in regression.

Chapter VIII

THE MV REGRESSION AND MV SIM COMPUTER PROGRAMS

The purpose of this chapter is to describe a computer program, called MV REGRESSION, which performs automatic regression analysis on a sample of size n. Also briefly outlined is program MV SIM which generates samples of size n from a specified p variate normal. The detailed operation of MV SIM is described in appendix A. Both programs are written in NELIAC compiler language. Operation of these programs on the Control Data Corporation model 1604 computer at the U. S. Naval Postgraduate School has produced all of the computations involved in the examples throughout this paper as well as the test results discussed in chapter IX and appendix B.

Briefly, MV SIM will analyze a specified p variate normal (given by U and $\sum$) and print out true regression coefficients and associated $\sigma_{pp.1,\ldots,q}$ for any set(s) of q variables specified by the program user ($q \leq p-1$). Next, MV SIM will generate a sample of size n from the specified p variate normal and compute sample vector Z and sample V-C matrix, S. Before turning control to program MV REGRESSION, MV SIM performs statistical tests on Z and S, and prints out results of these tests, but takes no action based on these results. These statistical tests and actual computer run results are discussed in detail in appendix B.

Before proceeding with a description of MV REGRESSION, it is interesting to consider the powerful research tool one has when he can specify a p variate normal (U and $\sum$) and quickly generate random samples from that distribution. It is obvious that this operation saves much time in gathering data, or in "making up" reasonable samples when it is desired to test the operation of a regression analysis program such as MV REGRESSION. (This was the case when computations for example in this paper were required). But MV SIM offers the statistician a much more useful research capability than this. Using MV SIM one can make accurate comparisons of the results of any regression scheme with true regression equations, conditional variances, etc., which MV SIM computes from the specified U and Σ. Of course, for such a comparison, the regression scheme must be applied to a sample drawn by MV SIM from the distribution specified by U and Σ.

The sampling capability of MV SIM also makes it possible to perform empirical sampling studies of random variables whose distributions are difficult to find theoretically. One such study, now in progress, is discussed in chapter IX.

We finish this chapter with a detailed description of program MV REGRESSION.

The inputs of MV REGRESSION are as follows:

1. Start with a sample of n observations of the p variate normal. If MV SIM supplies the sample, it will supply it in the form of Z and S.
2. Specify "standard" or "cost" option (see chapter VII). If cost option, give cost of observation, c_i, for variables y_i, for $i = 1,\ldots,p-1$. If the user specifies "standard", he still may specify costs and obtain

printed cost data even though the "regular" criteria is used as far as entering variables into regression is concerned.

3. Specify criteria for halting regression of a sample:

   A. 1) $F_{\alpha(1,n)}$, the value to be compared with statistic F for adding variables to regression.

      2) Min F, a value less than $F_{\alpha(1,n)}$ to be compared with statistic F for removing variables from regression.

   B. 1) Last variable added reduced the conditional variance of y_p by less than λ (not used for cost option).

      2) Last variable added, y_k, costs more than C dollars to observe per unit of variance reduction of y_p due to adding y_k (used only for cost option).

   C. Conditional variance of y_p became less than T.

   D. Number of variables in regression reached W.

Before step 1 of the regression operation, MV REGRESSION prints out s_{pp}, and (optionally) the RR matrix. (The RR matrix is a pxp matrix which contains all current data in compact form from which all required parameters at each step can be computed. Initially, it is a matrix of sample correlation coefficients which is easily computed from sample V-C matrix S. See Efroymson [B]).

At each step, after a variable has been added to regression, the following data is printed:

1 a. "Best" variable to have been added (variable with

minimum $\sigma_{pp.1,\ldots,q}$).

b. "Cheapest" variable to have been added.

c. Whichever of the two variables above that actually was added (a. if regular option, b. if cost option)

II The value used in the F test for the added (or removed) variable. MV REGRESSION compared this value with the input value of $F_{\alpha(1,n)}$ (or Min F).

III a. The square of the estimated new multiple correlation coefficient of y_p on the variables in regression.

b. The estimate of the new conditional variance of y_p, $s_{pp.1,\ldots,q}$.

IV The cost to observe the variable just added, y_k, per unit of conditional variance reduction due to the addition of this variable to regression <u>at this time</u>. This is computed as $c_k / (s_{pp.1,\ldots,q} - s_{pp.1,\ldots,q,k})$.

V a. A list of the new set of variables, y_i, $(i = 1,\ldots,q)$ in regression.

b. The estimated regression coefficients, b_i.

VI The cost to observe the new set of variables in regression per unit of total variance reduction of y_p. This is computed as:

$$\sum_{i=1}^{q} c_i \;/\; (s_{pp} - s_{pp.1,\ldots,q})$$

VII The new RR matrix (optional)

As indicated earlier, it is possible to specify cost of observations, c_i, even though the standard option is used. In this case, items IV and VI are still computed and printed, but of course, the "best" variable to add (item Ia) is still the one actually added.

Each step, at which a variable is being removed from regression, item I above becomes "the variable just removed", and items II, III, V, VI, and VII only are printed.

Minor changes to the program can be made to cause it to print out other data after each step, such as estimated variances of the estimated regression coefficients.

The next few pages show the actual program output of a regression analysis performed by MV REGRESSION on a sample of size 300 of a five variate normal. This sample was generated by the MV SIM program using input vector U and V-C matrix $\sum$ given by 4.1 and 4.2.

```
MULTIVARIATE ANALYSIS (CONTINUED)  2  19  1963     PAGE  7

COMPUTER RUN DATA
NUMBER OF SAMPLES = 3


CRITERIA FOR CHOOSING WHICH VARIABLE TO ADD TO
REGRESSION (AMONG THOSE PASSING F TEST)

   MAXIMUM REDUCTION OF THE CONDITIONAL VARIANCE OF Y 5

HOWEVER,
   THE FOLLOWING COSTS OF OBSERVATION ARE SPECIFIED

     Y1          Y2          Y3          Y4          Y5
   10.0000     12.0000     16.0000     20.0000      .0000



ANY ONE OF THE FOLLOWING CONDITIONS CAN HALT REGRESSION STEPS

   1) NUMBER OF VARIABLES IN REGRESSION REACHED 4
   2) CONDITIONAL VARIANCE OF Y 5 BECAME LESS THAN 4.0
   3) LAST VARIABLE ADDED REDUCED THE CONDITIONAL VARIANCE
        OF Y 5 BY LESS THAN  2.0
   4) LAST VARIABLE ADDED COSTS MORE THAN 10.00 DOLLARS
        TO OBSERVE PER UNIT OF VARIANCE REDUCTION OF Y 5
   5) NO MORE VARIABLES (AMONG THOSE NOT IN REGRESSION)
        PASS THE F TEST OF SIGNIFICANCE
```

```
MULTIVARIATE ANALYSIS (CONTINUED)  2  19  1963      PAGE  8
SAMPLE NUMBER  1
SAMPLE OF SIZE  3C0 OF THE 5 VARIATE NORMAL
SAMPLE MEANS
      Y1          Y2          Y3          Y4          Y5
    7.4125     48.2586     11.9077     29.8140     95.4458
SAMPLE VARIANCE CCVARIANCE MATRIX
      Y1          Y2          Y3          Y4          Y5
   31.5587     14.6327 -   28.5108 -   17.8985     55.3288
   14.6327    227.5360 -    3.9154 -  243.4449    171.7598
-  28.5108 -    3.9154     39.1032 -    7.6404 -   39.8334
-  17.8985 -  243.4449 -    7.6404    276.5990 -  191.4721
   55.3288    171.7598 -   39.8334 -  191.4721    199.0421
```

```
MULTIVARIATE ANALYSIS (CONTINUED)  2  19  1963      PAGE 10
ANALYSIS OF SAMPLE NUMBER  1

SAMPLE VARIANCE OF Y 5 =   199.0421
F LEVEL TO ENTER = 3.87      F LEVEL TO REMOVE = 3.7

RR MATRIX TO START
      Y1          Y2          Y3          Y4          Y5
    1.0000      .1726 -     .8116 -     .1915       .6981
     .1726     1.C000 -     .0415 -     .9703       .807C
-    .8116 -    .C415      1.0000 -     .0734 -     .4515
-    .1915 -    .9703 -     .0734      1.0CC0 -     .8160
     .6981      .8070 -     .4515 -     .8160      1.0000
```

```
MULTIVARIATE ANALYSIS (CONTINUED)  2  19  1963     PAGE 11
STEP    1
BEST VARIABLE TO ADD WAS Y  4
CHEAPEST VARIABLE TO ADD WAS Y  2
VARIABLE ADDED WAS Y  4

STATISTIC USED TO COMPARE WITH F(1,  298) =  595.9689
NEW MULTIPLE CORR COEFF SQUARED =      .3340
NEW CONDITIONAL VARIANCE =   66.7210
COST TO OBSERVE Y  4 IN DOLLARS PER UNIT VARIANCE REDUCTION =      .1511
NEW SET OF VARIABLES IN REGRESSION
     4
COEFFICIENTS     B(I)     B0 =  116.0842
-     .6922       .0000       .0000       .0000       .0000
COST TO OBSERVE THIS SET OF VARIABLES PER UNIT
   OF VARIANCE REDUCTION OF Y 5
      20.0000 DOLLARS DIVIDED BY
     132.3210 UNITS OF VARIANCE REDUCTION =      .1511
THE NEW RR MATRIX
     Y1          Y2          Y3          Y4          Y5
     .9632 -     .0132 -     .8256       .1915       .5417
-    .0132       .0583 -     .1128       .9703       .0152
-    .8256 -     .1128       .9946       .0734 -     .5114
-    .1915 -     .9703 -     .0734      1.0000 -     .8160
     .5417       .0152 -     .5114       .8160       .3340
```

```
MULTIVARIATE ANALYSIS (CONTINUED)  2  19  1963     PAGE 12
STEP   2
BEST VARIABLE TO ADD WAS Y  1
CHEAPEST VARIABLE TO ADD WAS Y  1
VARIABLE ADDED WAS Y  1

STATISTIC USED TO COMPARE WITH F(1,  297) = 3089.6640
NEW MULTIPLE CORR COEFF SQUARED =      .0293
NEW CONDITIONAL VARIANCE =    5.8889
COST TO OBSERVE Y  1 IN DOLLARS PER UNIT VARIANCE REDUCTION =      .1643
NEW SET OF VARIABLES IN REGRESSION
     1          4
COEFFICIENTS      B(I)     BO =  102.8895
    1.4124 -     .6008      .0000       .0000       .0000
COST TO OBSERVE THIS SET OF VARIABLES PER UNIT
   OF VARIANCE REDUCTION OF Y 5
      30.0000 DOLLARS DIVIDED BY
     193.1531 UNITS CF VARIANCE REDUCTION =     .1553
THE NEW RR MATRIX
     Y1          Y2          Y3          Y4          Y5
    1.0380 -     .C137 -     .8571       .1988       .5624
     .0137       .0581 -     .1241       .9730       .0226
     .8571 -     .1241       .2868       .2376 -     .0470
     .1988 -     .9730 -     .2376      1.0380 -     .7082
-    .5624       .C226 -     .0470       .7082       .0293
```

```
MULTIVARIATE ANALYSIS (CONTINUED)  2  19  1963     PAGE 13
STEP   3
BEST VARIABLE TO ADD WAS Y  2
CHEAPEST VARIABLE TO ADD WAS Y  2
VARIABLE ADDED WAS Y  2

STATISTIC USED TO COMPARE WITH F(1,  296) =  127.4672
NEW MULTIPLE CORR COEFF SQUARED =      .0205
NEW CONDITIONAL VARIANCE =    4.1344
COST TO OBSERVE Y  2 IN DOLLARS PER UNIT VARIANCE REDUCTION =    6.8394
NEW SET OF VARIABLES IN REGRESSION
     1          2          4
COEFFICIENTS      B(I)     B0 =   75.6177
    1.4258      .3643 -    .2792       .0000       .0000
COST TO OBSERVE THIS SET OF VARIABLES PER UNIT
   OF VARIANCE REDUCTION OF Y 5
        42.0000 DOLLARS DIVIDED BY
       194.9076 UNITS OF VARIANCE REDUCTION =      .2154
THE NEW RR MATRIX
     Y1          Y2          Y3          Y4          Y5
    1.0413      .2360 -     .8864       .4285       .5677
     .2360    17.1986 -    2.1349     16.7347       .3895
     .8864     2.1349       .0218      2.3150       .0012
     .4285    16.7347 -    2.3150     17.3214 -     .3292
-    .5677 -    .3895       .0012       .3292       .0205
   3) LAST VARIABLE ADDED REDUCED THE CONDITIONAL VARIANCE
         OF Y 5 BY LESS THAN  2.0
```

Chapter IX

CURRENT STUDIES AND PROPOSALS FOR FUTURE RESEARCH

In this chapter we discuss tests that have been started using programs MV SIM and MV REGRESSION. Also, plans for future research are proposed.

Some tests (described in Appendix B) of a large number of samples generated by MV SIM have been completed.

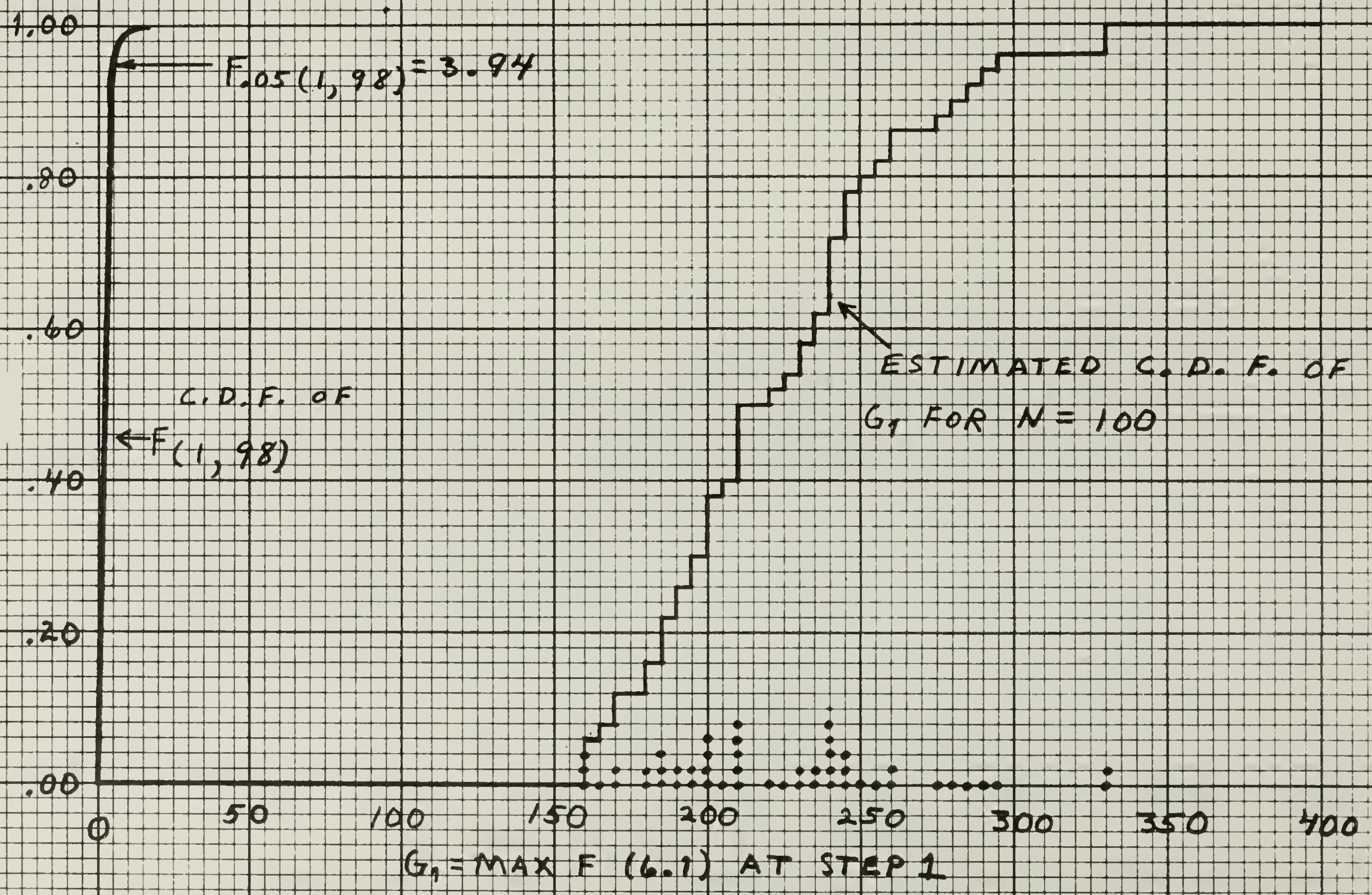
An empirical sampling study to study the random variables involved in the F test of chapter VI has been started since the form of the distribution is unknown and extremely difficult to obtain in closed form. Actually, p-1 random variables, which we will call $G_1,\ldots,G_{p-1}$, are under study at the same time. They are defined by a specified p variate normal, the size of each sample of the p variate normal, n, and the method of computing values of G_i, $i = 1,\ldots,p-1$, from a sample which is described next.

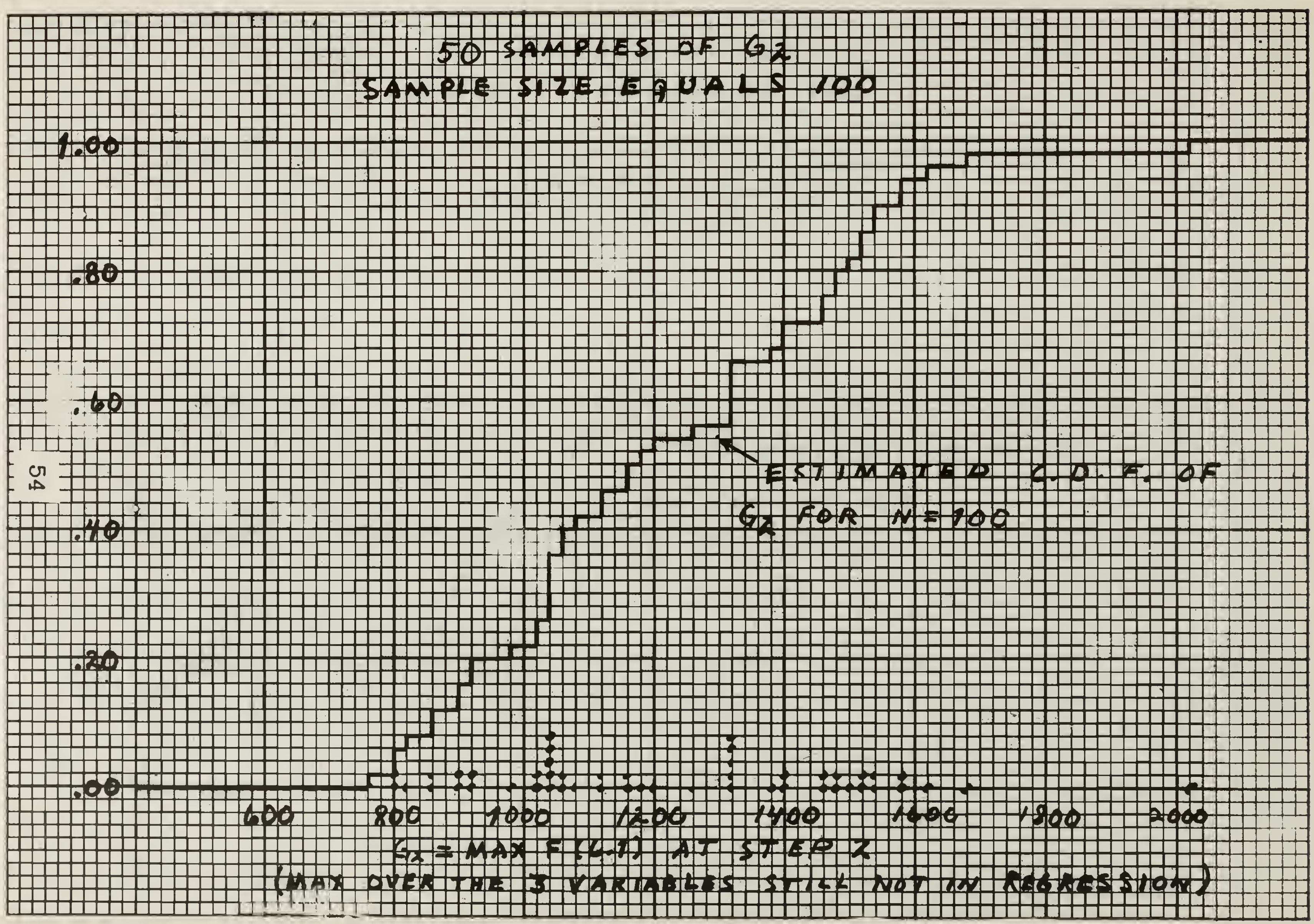
At step one, G_1 is defined as the maximum value of (6.1) where F is computed for each of the p-1 variables (none of which are in regression yet). G_2 is dependent upon G_1 in the sense that G_2 is the value of max $F(6.1)$ computed after the variable for which F equals G_1 has been entered into regression. Thus, at step two, max F is the maximum value of F for those p-2 variables still not in regression. The step-wise procedure continues without the use of any tests for halting so that a new variable is added at each step. Thus, at step i, G_i equals max F, where F is computed for each variable still not in regression by step i. After G_i is recorded, the variable for which $F = G_i$ is entered into regression.

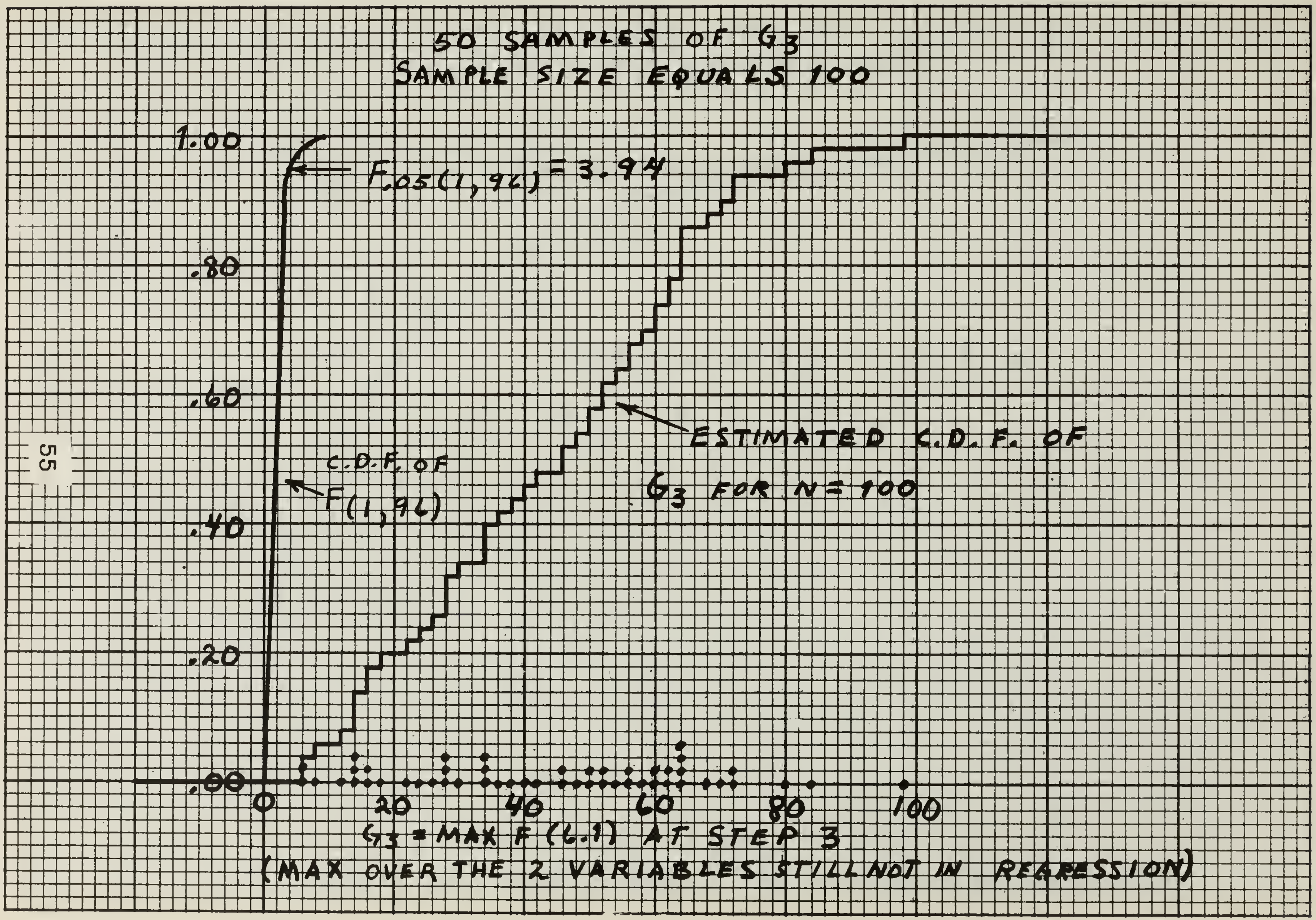
Since the values of $\left|6.1\right|$ depend upon the sample size, we see that each sample of size n of a specified p variate normal produces one value of each of the random variables $G_1, \ldots, G_{p-1}$. Also, to obtain repeated sets of values of the same random variables, the sample size must be kept constant.

The tests that have been completed were performed on the five variate normal specified by formulas 4.1 and 4.2. Six sample sizes: 50, 100, 150, 200, 250, and 300 have been computed 50 times each. The results of G_1, G_2, G_3, G_4 for the sample size 100 are plotted below in the form of estimated cumulative distribution functions (c.d.f.'s). Where feasible, the graphs also show the curve of the c.d.f. of $F_{(1,n-q-1)}$. (Recall that if the F test of chapter VI had been applied, each value of G_{q+1} would have been compared with $F_{\alpha(1,n-q-1)}$ at step q+1, for q = 0, 1, 2, 3).

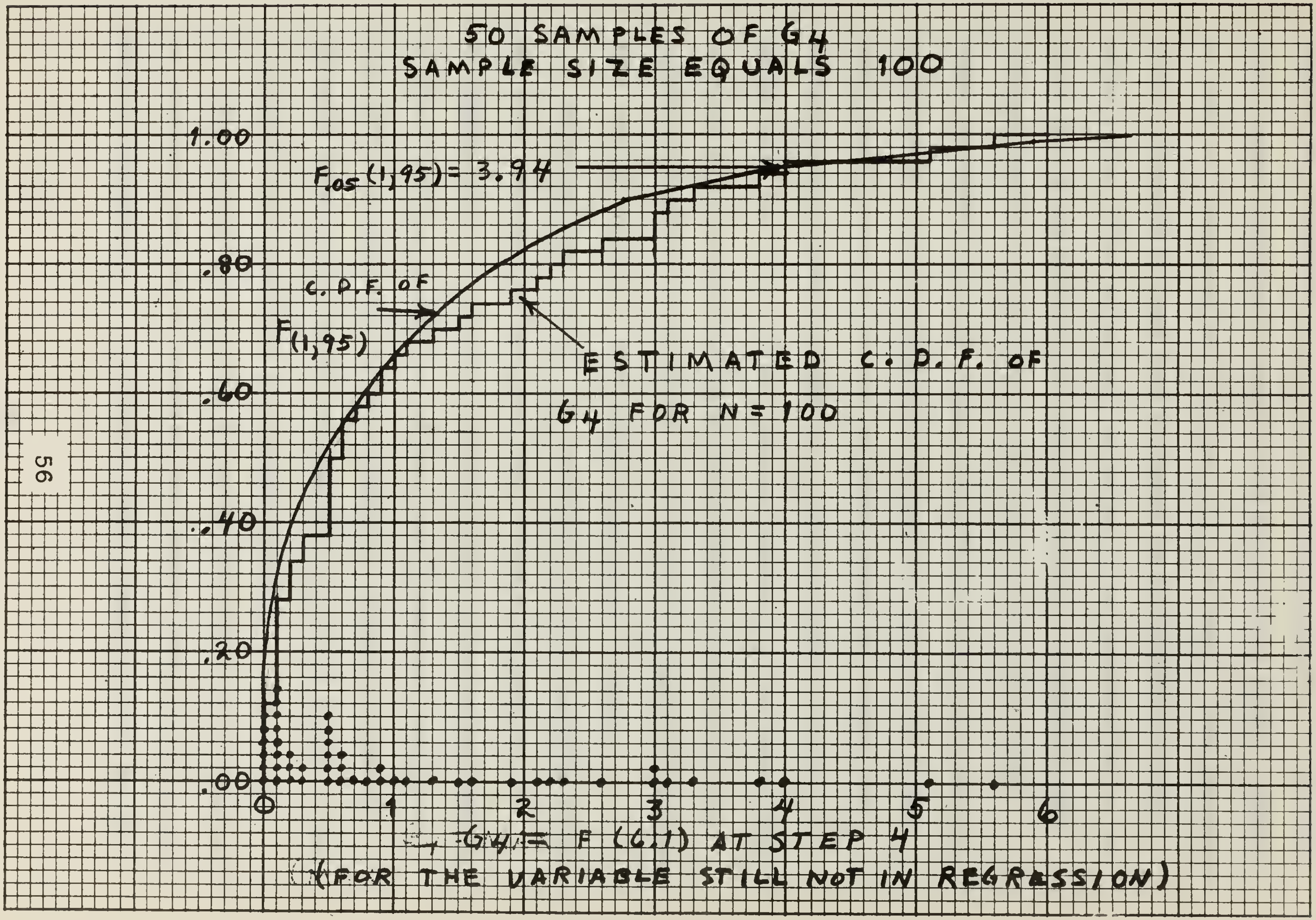
50 SAMPLES OF G_1
SAMPLE SIZE EQUALS 100




50 SAMPLES OF G_2
SAMPLE SIZE EQUALS 100
1.00
.80
.60
.40
.20
.00
ESTIMATED C.D.F. OF
G_2 FOR N=100
600
800
1000
1200
1400
1600
1800
2000
G_2 = MAX F(L,T) AT STEP 2
(MAX OVER THE 3 VARIABLES STILL NOT IN REGRESSION)



50 SAMPLES OF G_3
SAMPLE SIZE EQUALS 100


50 SAMPLES OF G4
SAMPLE SIZE EQUALS 100
1.00
.80
.60
.40
.20
.00
$F_{.05}(1,95) = 3.94$
C.D.F. OF
$F_{(1,95)}$
ESTIMATED C.D.F. OF
G4 FOR N=100
0
1
2
3
4
5
6
G4 = F (6,1) AT STEP 4
(FOR THE VARIABLE STILL NOT IN REGRESSION)

So far the min λ parameter test has not been implemented in program MV REGRESSION so the type of artifical control of the stepwise process described in chapter VII has not been tested. However, a number of samples of size 300 of an 18 x 18 matrix (the same matrix used in Appendix B) have been processed by MV REGRESSION, using rather wide limits on the halt criteria. After examination of the first run it was obvious that three variables in regression were too many and that either one or two would be the right number. Since the sample size was large, most samples allowed nine or more variables to enter regression on the basis of passing the F test even though nearly all of the variables beyond two reduced the estimated conditional variance of y_{18} by less than 1.0 unit. By comparison, the first variable usually reduced s_{18} from about 18.6 to about 6.5. An examination of the computed statistics of all variables (whether in regression or not) made it apparent that some test such as the min λ test might be quite useful here.

Advantage was taken of the fact that the true p variate normal was known when samples obtained from it were being analyzed by MV REGRESSION. For example, after the first run on several samples of the 18 variate normal, only six of the 17 possible predictors ever got into regression by the third step. Hence, all possible pairs of these six variables were fed back to MV SIM for which the true conditional variances of y_{18} were computed.

The various halt criteria suggested in chapter VII can be useful in developing methods of searching for optimal combinations of variables in regression. It is proposed that procedures, such as the one described below, be tested and compared with procedures already described to see if better results can be obtained.

We will assume that an experimenter has a large sample that perhaps was very expensive to obtain. We shall permit the experimenter two computer runs on the sample sample: the first run providing a set of feed-back data for the second run.

The main purpose of the first computer run is to determine a lower bound on the conditional variance of y_p. This is accomplished by using the F (and min F) test with the step-wise procedure with α set to permit most variables to enter regression. Of course, at each step valuable information such as the conditional variance of y_p, and the amounts of variance reduction due to each variable should be printed.

From the first run the experimenter chooses the maximum number of variables, say m, that he will have in his final prediction equation. This is usually easy to do by examining the decreasing values of $s_{pp.1}$, $s_{pp.1,2,...}$, $s_{pp.1,...,q}$; where q, q $\geq$ m, represents the number of variables in regression after the first computer run.

The purpose of the second computer run is to make a rather thorough (but not exhaustive) search for the optimal combination of m variables in regression. The procedure is to conduct p-1 separate regressions, each regression starting with a different first variable, and continuing until m variables are in regression. At each step (after the first), the variable chosen to enter regression will be the variable that can contribute most reduction in the conditional variance of y_p, unless, by adding this variable, a combination that had been in regression previously (during a previous regression) would result. For example, if the first regression added variables in order y_1, y_5, y_2, then if the second regression proceeded as y_2, y_5, variable y_1 would not be permitted to enter regression next. Instead, the second best variable would be chosen at this step.

Thus, after the second computer run is completed the experimenter will have (p-1)xm prediction equations (and conditional variances of y_p) to choose from, p-1 for each number of variables in regression.

Two further investigations are proposed. In Appendix B, the results of tests of a number of samples of a five and an 18 variate normal are described. As a result of the failure of the sample V-C matrices, S, of the 18 variate normal to pass the chi-square test, it is proposed that further testing of the multivariate normal generator be conducted. As indicated in Appendix B the possibility of round off error should be considered.

It is also suggested that a study be made to ascertain which of the two suggested tests of the matrix S is better. Possibly a study would indicate weakness in both. Anderson [1], section 10.8, describes a third test of matrix S.

The step-wise procedure of regression analysis as described in this paper is called the "forward" method because it starts with no variables in regression and adds them to regression one at a time. This is because the forward procedure permits computational shortcuts so that the number of computations can be minimized (especially so when Efroymson's computer program algorithm is used [3]). The backward operation of removing extraneous variables, however, offers no computational advantages. See Quenouille [8]. Another reason the forward procedure can be done with fewer computations is because, usually the number of variables in the final regression is much less than p-1. Often the reason for a large number of independent variables to be examined compared to the number finally used, is that from those variables actually measured additional variables are often created to account for possible curvilinearity and interaction. For example, if X_1 is a variable whose value was actually measured, variables $Y = X_1^2$, $Z = X_1^3$ may be computed and used as part of the original p-1 possible predictors. [9], see page 20.

One possible advantage in using the backward method is to start the process by computing an estimate of the lowest possible value of the conditional variance of y_p, $s_{pp.1,\ldots,p-1}$. If somehow this value could be obtained before the forward procedure was performed, one could estimate the amount of reduction available in the combined combination of variables still not in regression at each step. Knowledge of this value at each step should be useful in deciding which way would be best to go next: i.e., eliminate the weakest variables now in regression, or add the strongest variable still not in regression, or to halt.

BIBLIOGRAPHY

1. Anderson, T. W. "An Introduction to Multivariate Statistical Analysis" New York: John Wiley and Sons, Inc., 1960.

2. Barron, Joseph M. Thesis: "Random Number Generation on the CDC-1604," Monterey, Calif.: United States Naval Postgraduate School: 1962

3. Efroymson, M. A. "Multiple Regression Analysis" - (Ralston-Wilf, "Mathematical Methods for Digital Computers") New York: John Wiley and Sons, Inc., 1960.

4. Graybill, Franklin A. "An Introduction to Linear Statistical Models," Vol. I, New York: McGraw-Hill, Inc., 1961.

5. Green, Bert F. Jr., Smith, J. E. Keith, Klem, Laura. "Empirical Tests of an Additive Random Number Generator," Journal of the Association for Computing Machinery, Vol. 6, No. 4, October 1959.

6. Hald, A. "Statistical Theory with Engineering Applications," New York: John Wiley and Sons, Inc., 1952.

7. Miller, Robert G. "The Screening Procedure," Studies in Statistical Weather Prediction, Hartford, Conn.: Travelers Weather Research Center, December 1958.

8. Quenouille, M. H. "Note on the Elimination of Insignificant Variates in Discrimenatory Analysis," Annals of Eugenics, Vol. 14, pages 305-308. 1949.

9. Schultz, E. Fred and Goggans, James F. "A Systematic Procedure for Determining Potent Independent Variables in Multiple Regression and Discriminant Analysis." Agricultural Experiment Station, Auburn University, Auburn, Alabama, November 1961.

10. Vaa, Lcdr. Norman A. Thesis: "Generation and Testing of Random Numbers of an Arbitrary Distribution," Monterey, Calif.: United States Naval Postgraduate School: 1962

11. Wold, Herman, "Tracts for Computers," London: Cambridge University Press, 1948.

Appendix A

GENERATION OF THE P VARIATE NORMAL BY PROGRAM MV SIM

For the construction of each sample (of size one) from the specified p variate normal, MV SIM uses an independent sample of size p from the normal (0,1) distribution. (e.g., mean u = 0, variance σ = 1).

To obtain each independent normal random sample (of size one), MV SIM computes a function of an independent sample of size 12 from the uniform (0,1) distribution. (e.g., uniform on the interval zero to one). That this function only approximates normally distributed random numbers will be shown below.

It follows from the above that to generate a sample of size n of a p variate normal, nxpx12 random numbers from the uniform (0,1) random number generator are required.

A discussion of several techniques for generating uniformly distributed "pseudo" random numbers is given by Barron [2]. Empirical test procedures are also given.

The particular uniform (0,1) pseudo random number generator used by MV SIM is a subroutine called RAND. RAND was programmed according to specifications given by Green, Bert F. Jr., Smith, J. E., and Klem, Laura [5]. The number of initial random numbers, n in the reference, used by RAND is seven. This article also discusses a number of empirical tests that have been applied to this method.

The method by which MV SIM uses 12 independent uniform (0,1) random numbers to compute each pseudo normal (0,1) random number is discussed by Vaa [10], see page 40. Briefly, each normal random number is computed as:

$$X_i = \left(\sum_{j=1}^{12} W_j\right) - 6,$$

where the W_j are the required independent sample of size 12 from the uniform (0,1) distribution. The variance of the uniform (0,1) distribution is one-twelfth and variances of independent, uniformly distributed random variables are additive under convolution. Hence it is convenient to select 12 as the number of uniform random variables whose sum will approximate a normal variable. Means of (independent) uniform variables are also additive so that it remains to subtract the constant six from the sums of 12 independent uniform (0,1) random variables to approximate the normal (0,1) distribution. Vaa has a discussion of the advantages and disadvantages of this "truncated" approximation to the normal distribution.

Wold [11], pages xi to xiii, describes the method which MV SIM uses to convert an independent sample of size p from the normal (0,1) distribution to a sample from a p variate normal specified by U and $\sum$. This method requires the computation of a pxp triangular P matrix, $P = \{p_{ij}\}$, from the original V-C matrix, $\sum$, so that the following matrix equation holds:

$$\sum = P \cdot P^T$$

For our discussion we arbitrarily choose the triangulation of

$P = \{p_{ij}\}$ so that $p_{ij} = 0$ when $j > i$, i.e., let all "upper diagonal" elements of P equal zero. Next, assuming $x_1, \ldots, x_p$ is an independent sample of size p from normal (0,1), the sample of size one of the p variate normal is computed as:

$$\begin{aligned} y_1 &= u_1 + p_{11}\, x_1 \\ y_2 &= u_2 + p_{21}\, x_1 + p_{22}\, x_2 \\ &\vdots \\ y_p &= u_p + p_{p1}\, x_1 + \ldots + p_{pp}\, x_p, \end{aligned}$$

where the u_i are the elements of mean vector U.

The term "pseudo" random number is customarily given to numbers generated by arithmetic means, see Barron [2], pages 5, 6, of which the RAND subroutine is one.

It is now clear that the samples of size n of the p variate normal generated by MV SIM, are themselves pseudo random numbers, since they are merely arithmetic functions of uniform pseudo random numbers. Perhaps in this context, the operation of this part of MV SIM might have been called "simulation" of a p variate normal, rather than "generation". To carry this process one step further, sample mean vector Z, and V-C matrix S, being arithmetic functions of a sample of size n, are likewise pseudo random matrices. As in the case of the pseudo uniform and normal random numbers, it is desirable that some empirical tests be applied to these pairs of pseudo random matrices.

Appendix B describes some tests in detail: one for vector Z, and one for matrix S. These tests are (optionally) performed by

MV SIM on each sample, but MV SIM takes no corrective action except to print out the value of the computed statistics and an indication of the proper distribution to be compared with the statistics.

The Sequential Operation of Program MV SIM is as follows:

1. Print out input mean vector U, and V-C matrix Σ, and other miscellaneous data identifying the computer run.
2. Compute the P matrix from Σ as described above. Optionally, the P matrix may be printed out.
3. List the variance of y_p, σ_{pp}.
4. Compute the prediction equation for y_p, for each combination of variables, $y_1, \ldots, y_{p-1}$ that are specified by the program user as input. For each such regression the following data are printed:
   a) regression number
   b) q+1 variate normal, where q is the number of variables in regression
   c) multiple correlation coefficient (squared)
   d) conditional variance of y_p, $\sigma_{pp.1,\ldots,q}$
   e) the regression coefficients, β_i (optional)
5. Print out input data regarding samples of the specified distribution as described and illustrated in chapter VIII i.e., numbers of samples, observation costs, whether "standard" or "cost" option is used, etc.

The following operations are performed on each sample specified:

6. Generate the required sample of the specified p variate normal.

7. Compute samples mean vector Z, and V-C matrix, S. Print out Z and S.

8. Test sample means, Z (optional) (see Appendix B). Print out eigenvectors and eigenvalues of matrix, S, from which the proper statistic is computed. Also print out the statistic and the proper degrees of freedom of F to be used for comparison.

9. Test sample matrix, S (optional) (see Appendix B). Print out eigenvalues of sample matrix, S. Print out the statistic to be compared with chi-squared distribution. Also print out proper degrees of freedom to be used for comparison.

Of course, the user of program MV SIM can omit some of the above items such as items 3 and 4 at his discretion.

The actual analysis of each sample and associated printed output performed by MV REGRESSION is described and illustrated in detail in chapter IX.

Appendix B

TESTS OF SAMPLE MEAN VECTOR, Z, AND SAMPLE VARIANCE-COVARIANCE MATRIX, S

For a discussion of some of the problems encountered in generating random numbers by arithmetic means, see Barron [2] and Vaa [19].

Graybill [4], page 206, shows that if Y is a p variate normal with mean vector U and V-C matrix Σ, then the quantity:

$$v = (Z - U)^T S^{-1} (Z - U) (n - p) / p (n - 1),$$

is distributed as $F_{(p,n-p)}$, if indeed Z and S are computed from a sample of size n from the specified p variate normal. Hence, to test a sample mean vector, Z, an appropriate level, α, (usually .05) is chosen. Then if v is less than $F_{\alpha(p,n-p)}$, vector Z is accepted as having been computed from a reasonable sample; otherwise Z is rejected.

To perform a test for a sample V-C matrix, S, an orthogonal transformation is performed on both Σ and S, separately, yielding diagonal matrices Δ and D respectively. Δ is a V-C matrix of a p variate normal with independent variables (i.e., all covariances are equal to zero). Now, if it is true that S is computed from a sample drawn from a p variate normal with V-C matrix, Σ, then D should be a sample drawn from a p variate normal with V-C matrix, Δ. Hence, a test that D is a sample from Δ should verify that S is a sample from Σ.

Since each element of D, s'_{ii} $(i = 1,\ldots,p)$, is a sample variance, and since each element of Δ, σ'_{ii}, is the true variance corresponding to element s'_{ii}, for all i, intuitively, it appears that each of the statistics:

$$(s'_{ii} / \sigma'_{ii}) \cdot (n - 1) \quad i = 1,\ldots,p ,$$

should have the chi-square distribution with n - 1 degrees of freedom (n is still the sample size). From this, and the fact that s'_{ii} is statistically independent of s'_{jj} for all i, j = 1,...,p (i ≠ j), it follows that the statistic:

$$\text{(B1)} \qquad (n - 1) \sum_{i=1}^{p} (s'_{ii} / \sigma'_{ii})$$

has the chi-square distribution with p·(n-1) degrees of freedom, since the degrees of freedom of sums of independent chi-squares are additive.

Hence, to test each sample V-C matrix, S, MV SIM "rotates" Σ and S, and computes formula B1 above from Δ and D. Printed out (optionally) are the p diagonal elements of Δ and D (the eigenvalues of matrices Σ and S respectively). Also printed are the result of formula B1 and the number of degrees of freedom of the chi-square distribution to be used for comparison.

Programs MV SIM and MV REGRESSION were used to generate and test a number of samples from two different p variate normals. One of these normals is specified by 4.1 and 4.2 (five variate normal). The other distribution was an 18 variate normal that was very close to being singular. (Several sets of rows were close to each other in value).

Six sample sizes: 50, 100, 150, 200, 250, and 300 were studied of the five variate normal, with 20 samples tested of each size.

Four sample sizes: 50, 100, 150, and 200 were studied of the 18 variate normal, with 20 samples tested of each size.

For the five variate normal, both statistics (for Z and S) appeared to behave as samples from their respective F and chi-squared distributions for all sample sizes.

However, curious results were obtained from the unusual 18 variate normal tested. All Z tests passed as nicely as for the five variate normal. However, the values of chi-square were much too high, indicating poor sample V-C matrices, S, were being generated. For example, for the 20 samples of size 100 (of the 18 variate normal) the statistic BI should behave as chi-square with 1782 degrees of freedom (which is the mean of that distribution). The 20 computed values of BI ranged from 2213 to 2683.

A possible reason for these poor results could be due to the use of a poor random number generator. However, the satisfactory results obtained from testing the five variate normal, as well as tests of the uniform random number generator conducted previously leads one to seek a different source of error.

Possibly a more reasonable explanation is the likelihood of computer round off error. The large number of computations required to rotate an 18 x 18 matrix plus the fact that the matrices were all nearly singular could very likely cause this type error. If this is the case, the generated sample V-C matrices themselves may be "good" samples that are merely difficult to test.

Another interesting possibility is the method used to rotate matrix S for the test. Recall that rotating a symmetric matrix, $\sum$, to yield a diagonal matrix, Δ, can always be done by finding an orthogonal matrix, R_1, so that the following is satisfied:

(B2) $$R_1^T \cdot \Sigma \cdot R_1 = \Delta .$$

Also since S is also symmetric R_2 can be found so that

$$R_2^T \cdot S \cdot R_2 = D ,$$

where D is diagonal. Since Σ and S are not exactly equal it follows that orthogonal matrices R_1 and R_2 will not be equal.

Perhaps one might argue that a "better" test might be to find R_1 from the rotation of Σ, B2 above, and then compute:

$$R_1^T \cdot S \cdot R_1 = D'$$

where D' should be nearly diagonal if S is a reasonable sample from Σ; then compare the diagonal elements of D' and Δ as described above for D and Δ.